图书情报工作
第64卷 第3期 2020年2月

基于关键词关联度指标(KRI)进行LDA噪声主题过滤的方法研究*

■ 蒋甜　刘小平　刘会洲
中国科学院文献情报中心　北京100190

摘　要：[目的/意义]针对LDA模型主题识别结果通常包含噪声主题的问题，建立科学有效的主题过滤方法，排除噪声主题，确保主题识别及后续演化分析的准确性。[方法/过程]基于关键词之间的共现关系，构建关键词关联度指标(KRI)，借助定量手段进行主题筛选和过滤。以单细胞研究领域为例，计算各主题－关键词分布的KRI值，与人工判读结果进行对比分析。[结果/结论]实验结果表明，该方法能够有效排除LDA模型识别结果中的噪声主题，提高主题识别的准确性，也在一定程度上降低了主题识别过程对人工判读的依赖性。

关键词：主题过滤　LDA模型　关键词关联度指标KRI
分类号：TP393
DOI：10.13266/j.issn.0252－3116.2020.03.010

1　引言

科技文献是科学技术发展过程中重要的知识载体，蕴涵丰富的有情报价值的主题内容。近年来，很多研究者尝试采用不同的方法对海量的文献知识库进行信息分析和文本主题识别，以辅助科研人员快速把握文本主题，追踪科技领域主题演化规律，提高科研效率。主题模型方法能够从语义层面深入挖掘“文本－主题－词语”之间的隐含关系，是学科主题演化的重要研究方法。LDA模型即潜在的狄利克雷分布模型(Latent Dirichlet Allocation Model)，是一种经典有效的概率生成模型，包含文本－主题－词三层贝叶斯结构，可以从大规模文档集中挖掘潜在的主题信息[1]。LDA模型广泛应用于机器学习、信息检索、生物识别等多种领域，尤其在科技文献主题识别与演化研究中发挥着重要作用。

在利用主题模型进行学科主题演化分析的过程中，主题识别的精确性是基础，决定了后续步骤中构建的主题演化路径是否准确。LDA模型主题识别结果往往包含少数无效主题，这些主题经过人工判读难以解读出确切含义，对演化分析的精确性造成严重干扰，需要进行主题过滤。本研究基于LDA模型识别结果的主题－关键词分布中多词共现关系的统计分析，构建了关键词关联度指标(Keywords relevance index，KRI)，并以此为依据对主题识别结果进行筛选和过滤，去除无确切含义的噪声主题，避免了对主题演化研究的干扰。

2　相关研究

LDA模型自提出以来受到广泛关注和不断改进，在此基础上产生了经典的Dynamic Topic Model(简称DTM)模型[2]、Topic Over Time(简称TOT)模型[3]等，以及近几年用于微博等短文本分析的Biterm Topic Modeling(简称BTM)模型[4]、Hashtag-LDA模型[5]，可处理多类分类的有监督的Diagonal Orthant Latent Dirichlet Allocation(简称DOLDA)模型[6]等。

除了对传统模型的改进，模型质量的优化提升也是研究者关注的重点。作为一种无监督的机器学习方法，LDA模型生成主题的质量不尽如人意，有些主题无法解析出具体的含义，称之为噪声主题。噪声主题的存在直接影响LDA模型对文本数据的释义情况，因此

* 本文系中国科学院文献情报能力建设专项“科技领域战略情报研究与决策咨询体系建设”子课题“基础交叉前沿领域战略情报研究与决策咨询”(项目编号：Y8C0381005－01)研究成果之一。
作者简介：蒋甜(ORCID：0000－0002－9065－1223)，博士后，E-mail：jiangtian@mail.las.ac.cn；刘小平(ORCID：0000－0002－3342－8041)，研究员，硕士生导师；刘会洲(ORCID：0000－0002－7808－8570)，研究员，博士生导师。
收稿日期：2019－04－16　修回日期：2019－07－23　本文起止页码：92－99　本文责任编辑：杜杏叶

蒋甜，刘小平，刘会洲. 基于关键词关联度指标(KRI)进行 LDA 噪声主题过滤的方法研究[J]. 图书情报工作,2020,64(3):92－99.

非常有必要对 LDA 模型识别结果中的噪声主题进行过滤。目前主要的研究方法有以下几种：

(1)主题词判定法。这种方法的主要思想是认定在当前语料库中频繁出现而在一般英语中不常出现的词汇是主题词，认定非主题词为噪声词汇从而将其排除。谢琰[7]等利用一个外部语料库(Wikipedia 2014)生成词向量，根据词向量来计算两个单词的语义相似度，再与主题一致性中的同文档词频矩阵相结合，实现外部语料库对主题一致性的指导作用，从而更加精确地对主题质量进行评价，再通过设定阈值来过滤噪声主题，以提高主题模型的质量。

(2)主题概率分布法。曲佳彬等提出通过计算主题在所有文献中出现的概率，过滤掉在所有文献中出现概率低的主题[8]。这种方法有一个假设前提，即认为只有在所有文献中出现概率均较高的主题，才是反映某个时间段内文献主要内容的核心主题，对于分析主题演化有重要意义；反之，那些在多数文献中出现概率较低的主题，则很可能是边缘化甚至无意义的主题，对分析学科主题演化作用不仅不大，而且有可能干扰学科主题演化的分析效果。然而，某一时间区间内的新兴主题以及某些趋于衰亡的主题在所有文献中出现的概率并不高，通过这种方法很容易将这些类型的主题过滤掉，显然会对主题演化的准确性和科学性造成影响，不能真实客观地反应主题演化情况。

(3)基于信息熵的过滤方法。这是目前比较常用的主题过滤方法。词语在该主题下的概率分布越平均，主题的信息熵就会越大，通过设置主题信息熵的阈值可以达到过滤语义宽泛主题的目的。根据 LDA 模型输出的“主题－词汇”分布，可计算出每个主题的信息熵，其计算如公式(1)所示[9]：

$$Entropy(T) = -K\sum_{j=1}^{m} P_j \ln(P_j) \quad \text{公式(1)}$$

其中，K 为常数，P_j 表示主题 T 中第 j 个词的出现概率，该主题中共包含 m 个词汇。这种方法能够在一定程度上排除无效主题，但也具有较大的局限性。一是主题信息熵阈值的确定主观性太强，二是对于那些主题－关键词分布具有倾向性，但人工判读无法解析出确切含义的主题不能有效过滤。

(4)基于“垃圾主题”的过滤方法。李保利等提出通过计算由 LDA 模型产生的主题与定义的不能凸显文档内容的“垃圾主题”之间的相似度来进行主题过滤[10]。相似度越小说明该主题越能凸显文档的内容，设置合适的阈值过滤掉相似度较大的主题。“垃圾主题”可从“主题－词”的角度或“文档－主题”的角度定义。

(5)启发式方法。Y. L. Chang 等提出利用 Spike-and-Slab 先验分布基于文档来进行特征提取[11－12]，属于 slab 分布的词作为特征被保留进行主题估计，属于 spike 分布的词被过滤掉，提高了模型的可解释性且稀疏性较好，但该方法对于主题语义的抽取缺乏指导性原则。

综上所述，目前关于主题过滤的方法存在各自的局限性，过滤效果并不十分理想，特别是对于新兴主题、衰亡主题等文档数目较少的主题类型容易被当成噪声主题过滤掉。因此，探究新的主题过滤方法进一步提高主题过滤的精确性是非常必要的。本研究构建关键词关联度指标(KRI)进行主题过滤，对文档中多个关键词的共现频率进行统计分析，通过对不同共现词数赋予不同的权重，强化了多关键词共现在主题语义揭示中的“贡献率”。

3 基于 KRI 进行 LDA 噪声主题过滤的方法研究

3.1 基于 LDA 的主题识别

在利用 LDA 模型进行主题识别的过程中，主题数目的确定直接影响主题识别的效果[13－14]。主题数目设置过多，会造成识别出的主题分布过于稀疏，主题相似度过高；主题数目设置过少，会导致主题过于宽泛，无法准确揭示文献核心内容。在本研究中采用主题平均相似度和困惑度相结合的方法确定最优主题数目。

困惑度是用来评估语言模型优劣的指标，其基本评价方式是对测试集赋予高概率值的模型更好[15]。LDA 模型的困惑度计算公式(2)如下：

$$\text{Perplexity}(D) = exp\left\{ -\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\}$$

公式(2)

其中，D 表示语料库的测试集，M 为测试集中的文档数，N_d 表示第 d 篇文档的词汇数，$p(w_d)$ 表示第 d 篇文档中词汇的概率分布。

主题平均相似度是衡量所有主题之间平均差异程度的指标，通常基于 Jensen-Shannon 散度(JS 散度)来衡量[16]，计算公式(3)如下：

$$avg_sim(T_i, T_j) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} JS(T_i || T_j)}{K \times (K-1)/2}$$

公式(3)

其中，T_i 和 T_j 分别表示两个主题，JS($T_i || T_j$)表示主题 T_i 和 T_j 之间的 JS 散度。

从模型泛化能力的角度出发，困惑度越低，LDA 模

型的泛化能力越强[14]；从主题抽取效果考虑，主题平均相似度越小，说明主题之间的差异越大，重复主题越少，对应的 LDA 主题识别效果越好[17]。通常情况下，随着主题数目的增多，主题平均相似度增大，困惑度大致呈下降趋势，但会出现一些明显的拐点，这些拐点反映出在该主题个数时模型的泛化能力明显增强[8]。本研究综合考虑了困惑度和主题平均相似度，选取困惑度曲线下降趋缓处的拐点，比较主题平均相似度大小，结合实际主题识别效果，选定最优主题数目。

3.2 KRI 的构建

要想得到好的主题过滤效果，就需要一个客观的主题评价指标作为主题筛选和过滤的依据。在文献情报研究中，通常认为存在共现关系的词语能够揭示同一主题含义，从而基于共词分析进行文本挖掘[18-19]。在本研究中，通过对主题-关键词分布中，关键词之间的共现关系进行统计分析，筛选有效主题，过滤噪声主题。传统的共词分析往往只考虑词语两两之间的共现关系，某一主题下关键词两两共现频率（co-occurrence frequency）计算公式(4)为：

$$coof(W_a, W_b) = \frac{n_2}{N_2} \qquad \text{公式(4)}$$

其中，W_a 代表关键词 a，W_b 表示关键词 b，n 为同时包含关键词 a 和 b 的文档数目，N 为包含关键词 a 或 b 的所有文档数目。

除了两两共现的情况，关键词分布同时存在三词共现、四词共现等高阶共现的情况，类似两两共现频率计算公式，可以得到，三词共现频率计算公式(5)为：

$$coof(W_a, W_b, W_c) = \frac{n_3}{N_3} \qquad \text{公式(5)}$$

四词共现频率计算公式(6)为：

$$coof(W_a, W_b, W_c, W_d) = \frac{n_4}{N_4} \qquad \text{公式(6)}$$

以此类推，可以得到所有高阶共现的频率计算公式。

一个主题-关键词分布中，共同出现于同一篇文献的关键词数越多，同时包含这些关键词的文章数目越多，表明这个主题越“集中”，主题揭示度越高，该主题揭示的含义越准确。同时，高阶共现现象比低阶共现现象揭示更高的主题集中度，为突出多词共现对揭示主题含义的“贡献率”，将采用关键词共现数目的平方值作为权重。基于以上讨论，构建主题关键词关联性指标（Keywords relevance index，即 KRI），计算公式(7)如下：

$$KRI = 2^2 \sum coof(W_a, W_b) + 3^2 \sum coof(W_a, W_b, W_c) + \cdots + n^2 \sum coof(W_a, W_b, W_c, \cdots W_n) \qquad \text{公式(7)}$$

KRI 反映了主题中关键词共现的强度，揭示了该主题的关键词在不同文章中的分布集中度，为噪声主题的识别提供了量化手段。

4 实证研究

4.1 数据集构建

单细胞研究是生命科学领域的研究热点，是生命科学、材料科学、化学等多学科融合的交叉科学，单细胞技术广泛应用于胚胎植入前遗传学诊断[20]、干细胞与再生医学[21-22]、癌症诊断和治疗[23]、环境监测[24]等诸多方面，涉及的细分领域较多，要对其进行主题分析，对主题识别方法提出了较高的要求。为验证基于 KRI 进行主题过滤方法的有效性，以单细胞领域为例进行 LDA 主题识别及主题过滤。从 Web of Science 核心合集中检索得到 1990-2018 年单细胞领域相关文献 54 848 篇，类型为 Review、Article、Proceeding Paper 和 Letter，从中抽取每篇论文的标题、摘要和作者关键词作为主题识别和分析的语料。用 python 语言编写程序，调用 NLTK 库进行分词、词性标注、词干化、词形还原及去停用词等文本预处理。

4.2 基于 LDA 的主题识别及人工判读

采用 LDA 模型对构建好的数据集进行主题识别，计算主题数目 K 为 5-100，步长为 5 的困惑度和主题平均相似度，绘制困惑度-主题平均相似度曲线见图 1。随着主题数目不断增加，困惑度呈下降趋势，主题数目达到 30 之后，下降程度趋缓，表明 K=30 时模型的泛化能力增强[8]，主题数目达到 45 之后不再下降。综合考虑困惑度和主题平均相似度的值，选取 30 个主题的 LDA 模型输出结果，通过分析主题-关键词分布中概率较高的关键词及各个关键词之间的语义关系进行人工判读。

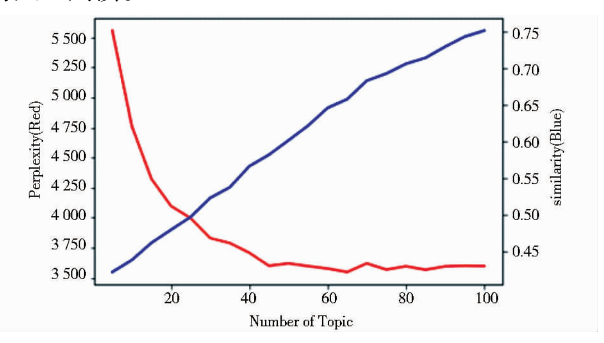


图 1 不同 K 值下 LDA 模型困惑度-主题平均相似度曲线

蒋甜，刘小平，刘会洲.基于关键词关联度指标(KRI)进行LDA噪声主题过滤的方法研究[J].图书情报工作,2020,64(3):92－99.

表1列出了采用LDA模型识别出的30个主题中的部分主题。针对每个主题,只展示在主题中出现概率较高的前10个词汇。从表1可以看出,有些主题的关键词能够较好地揭示主题内容,如主题13所含大多数词语与“基因表达调控”相关,主题23所含词汇均与“微生物燃料电池相关”。但并非所有的主题都能通过关键词解析出确切的主题内涵,如主题7中,其所含的高概率词“comparison”“datum”“reaction”“extent”“degree”都是含义非常宽泛的词汇,不能表征具体的含义,因此需要通过主题过滤排除这部分没有确切含义的噪声主题。

表1　K=30时LDA主题识别结果(部分主题)

主题序号	主题	主题词汇
3	干细胞培养	mouse、vivo、vitro、culture、differentiation、proliferation、stem、factor、survival、bone marrow
13	基因表达调控	expression、protein、gene expression、gene、mRNA、transcription、situ hybridization、regulation、promoter、bioenergy
23	微生物燃料电池	microbial fuel、cell、density、generation、removal、maximum power、wastewater、electricity、mW/m、power density
24	单细胞活体成像	detection、microscopy、imaging、sensitivity、fluorescence、measurement、living、quantification、sample、resolution
7	/	comparison、datum、situ、reaction、variation、single cell、extent、form、period、degree
12	/	contrast、betum、target、hypothesis、basal、addition、fusion、column、injury、majority

4.3　利用KRI进行主题过滤

根据3.2中构建的KRI对K=30时的LDA模型识别结果进行主题过滤,计算各主题－关键词分布的KRI值汇总并降序排列(表2),KRI值较高(>100)的主题用(H)标注,KRI值较低(<20)的主题用(L)标注。

由表2中看出,经人工判读无法解析出确切含义的无效主题,其KRI值均较低,在本实例中均低于20,说明KRI指数可以起到良好的主题过滤作用。

4.4　与词共现聚类方法的对比分析

KRI主题过滤方法借鉴了共词分析的思想,但又与传统的共词分析不同。共词分析法认为,当两个关键词或主题词在一篇文献中同时出现时,表示二者之间具有一定相关性,共同出现次数越多,相关性越大[25]。这种方法仅考虑了关键词两两之间的共现,忽视了实际存在的多个关键词共现情况,实际上多个关键词的共现更能体现主题的集中性和有效性。KRI方法采用多词共现分析,通过对多词共现频率的统计分析计算KRI值。在此也利用词共现聚类的方法做了主题分析,以便与经KRI过滤的LDA主题识别结果进行对比。

此处使用了可视化工具VOSviewer软件,利用共词分析得到词的共现网络见图2。共词分析法的结果得到了六个大主题,分别是基因表达调控(红色)、神经调节及钙调控(绿色)、微生物燃料电池(蓝色)、单细胞培养及分析(黄色)、单细胞动力学建模(紫色)以及单细胞凝胶电泳(天蓝色)。对关键词共现网络进一步分析,根据每个大主题内部的关键词簇,以节点关键词为核心又可以划分为若干子主题。如基因表达调

表2　K=30时LDA主题识别结果按KRI指数降序排列

主题编号	KRI指数	主题内容
23	10 468.970 25(H)	微生物燃料电池
21	1 624.146 247(H)	单细胞凝胶电泳
29	757.900 101 7(H)	单细胞油脂
19	496.705 424 9(H)	胚胎植入前遗传学诊断
28	294.724 320 8(H)	胞内钙调控
3	291.985 213 1(H)	干细胞体外培养
6	161.959 113 3(H)	运动的神经调节
13	152.734 512(H)	基因及蛋白表达调控
18	131.249 716 9(H)	肿瘤异质性
14	130.450 775 2(H)	免疫反应
0	96.627 829	胰岛素分泌及转运
2	87.672 684 46	肿瘤诊断和治疗
24	84.359 465 74	单细胞活体成像
17	69.713 891 82	细胞周期
10	69.388 108 52	单细胞蛋白
20	52.131 314 74	单细胞全基因组测序
5	50.105 380 24	细胞粘附
1	45.353 136 31	胚胎发生
11	39.321 852 06	胞外电子传递
15	34.559 458 65	细胞迁移
27	31.564 428 77	细胞形态学观察
9	26.790 711 12	单细胞数学建模
8	23.621 107 04	单细胞给药
26	23.545 202 22	单细胞质谱
16	19.792 553 73(L)	无效主题
22	18.384 718 5(L)	无效主题
12	15.536 748 93(L)	无效主题
25	15.222 222 22(L)	无效主题
4	12.328 025 48(L)	无效主题
7	12.074 777 76(L)	无效主题

控可以划分为干细胞基因表达调控、肿瘤细胞异质性、癌症诊断和治疗、单细胞原位杂交、流式细胞术 5 个子主题，神经调节及钙调控包含神经调节、胞内钙调控、受体激活、细胞凋亡 4 个子主题，微生物燃料电池包含微生物燃料电池产电性能评价、污水处理、生物修复、产电微生物胞外电子传递机制 4 个子主题，单细胞培养及分析包含单细胞培养、单细胞活体成像、微流控芯片、细胞迁移 4 个子主题，单细胞动力学建模包含单细胞生长动力学数学建模和单细胞蛋白 2 个子主题，单细胞凝胶电泳无细分子主题。

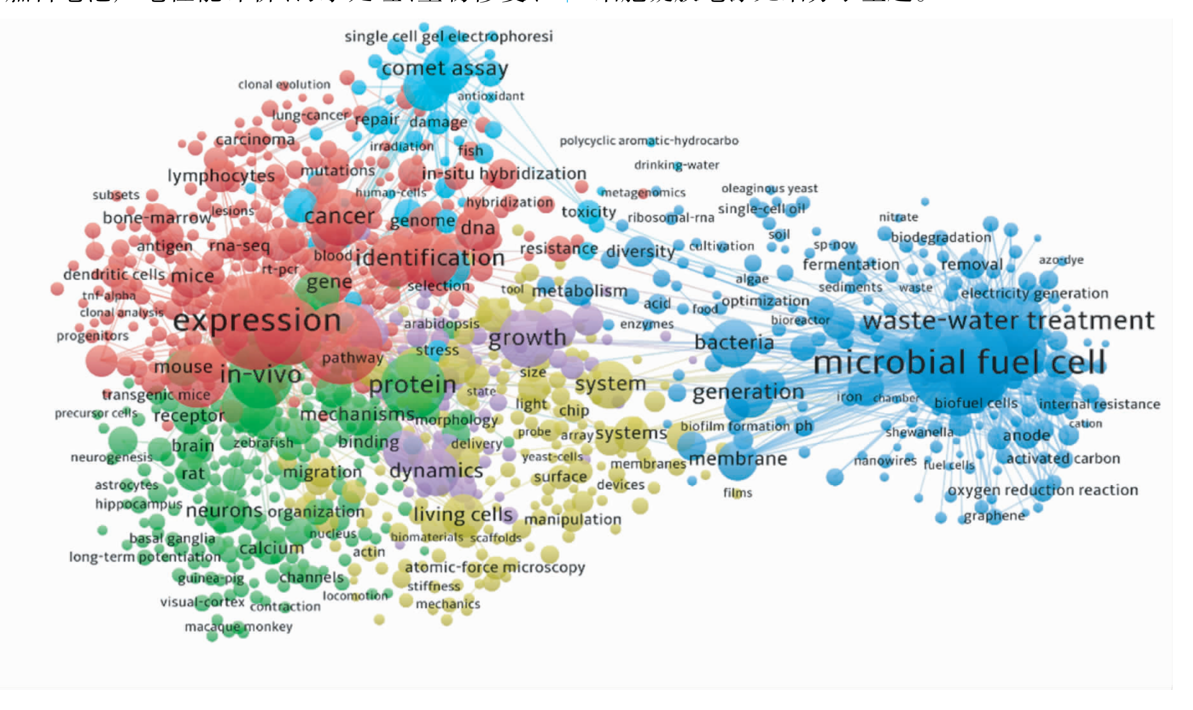


图 2　VOSviewer 软件对单细胞领域共词分析结果

通过以上的分析可以看出，共词分析可以准确识别出单细胞领域几个大的研究方向并将其划分为几个大主题，每个大主题下又可以根据节点关键词划分若干子主题，关键词共现网络还可以清晰地呈现出各个主题之间的层次关系及相互关联。然而，共词分析存在如下问题。第一，共词分析中存在许多孤立词，这些词和其他关键词之间缺乏关联，对于主题解析造成一定影响，如基于 KRI 过滤的 LDA 主题识别结果中的“单细胞油脂”“胚胎植入前遗传学诊断”“免疫反应”“细胞周期”“单细胞全基因组测序”“胚胎发生”等主题的核心关键词在 VOSviewer 关键词共现网络中均为孤立词，导致这些主题无法通过共词网络识别出来。第二，共词分析法受限于词频的影响，对于常规主题、热点主题的识别较容易，难以发现一些较小的、边缘的主题，不能保证主题的全面性和完整性。第三，共词分析对于主题颗粒度的把控有一定困难。LDA 模型中可以对主题数目的选择调整识别出的主题颗粒度，共词分析的结果易受共现频率阈值设定的影响，阈值过低，出现的关键词太多，对分析造成干扰，阈值太高，一些非热点主题又难以识别。第四，科技文献文本挖掘的目的通常并不限于主题识别，更关注主题演化情况。LDA 主题模型有先离散、后离散以及将时间信息结合到 LDA 模型中三种主要的主题演化分析思路，可以通过计算主题强度和主题相似度研究主题强度和内容的变化，便于定量地研究主题演化情况。相比之下，共词分析对时间元素的应用比较简单，不利于反映学科领域的更细化的发展和演化。

4.5　与“基于主题分布的边缘主题识别与过滤”方法的对比分析

对 K＝30 时的文档－主题分布进行统计分析，每个主题包含的文档数目与数据集文档总数的比值即为主题概率。由图 3 可以看出，主题的 KRI 指数曲线与文档－主题概率曲线二者趋势大致相同，特别是对于高概率主题的判断，两条曲线的重合度较高，但对于低概率主题的判断并不完全一致，例如主题 19 虽然主题概率值较低，但其 KRI 值在有效范围内，且人工判读为有确切含义的主题。

将 lgKRI 值及主题概率按照降序排列得到表 3。发现有些概率较高的主题并没有确切含义，相反有些低概率主题为有效主题。综合比较人工判读结果、KRI

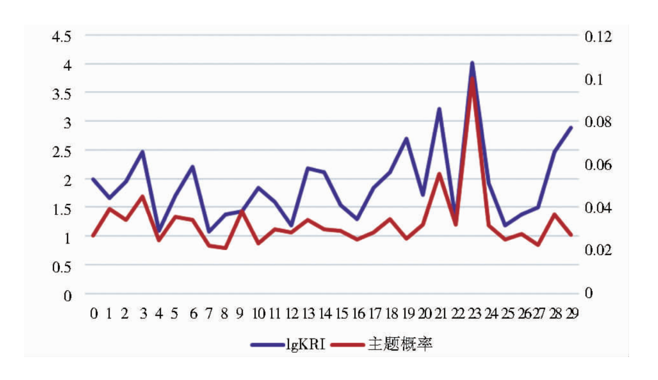


图 3 K＝30 时 LDA 识别结果 KRI 指数及文档－主题概率分布曲线

指数以及主题概率，KRI 指标的主题过滤效果优于“基于主题分布的边缘主题识别与过滤”的方法。

表 3 K＝30 时 LDA 主题识别结果 KRI 指数与主题概率按降序排列对比分析

主题编号	lgKRI	主题编号	主题概率
23	4.019 903 966(H)	23	0.099 711 931
21	3.210 625 133(H)	21	0.055 553 53
29	2.879 611 965(H)	3	0.045 051 779
19	2.696 098 903(H)	1	0.039 399 796
28	2.469 415 976(H)	9	0.038 141 774
3	2.465 360 858(H)	28	0.036 847 287
6	2.209 405 39(H)	5	0.035 753 355
13	2.183 937 182(H)	18	0.034 622 958
18	2.118 098 375(H)	13	0.034 240 082
14	2.115 446 664(H)	6	0.034 148 921
0	1.985 102 222	2	0.034 130 688
2	1.942 864 304	20	0.032 033 985
24	1.926 133 821	22	0.031 924 592
17	1.843 319 328	24	0.031 541 715
10	1.841 285 049	11	0.029 846 12
20	1.717 098 678	14	0.029 773 191
5	1.699 884 362	15	0.029 080 368
1	1.656 607 325	12	0.028 460 473
11	1.594 633 965	17	0.028 442 24
15	1.538 566 931	26	0.027 494 166
27	1.499 197 934	29	0.027 275 379
9	1.427 984 241	0	0.026 983 664
8	1.373 300 248	19	0.025 488 623
26	1.371 902 425	16	0.025 233 372
16	1.296 501 833(L)	25	0.024 996 354
22	1.264 456 984(L)	4	0.024 631 709
12	1.191 360 148(L)	10	0.023 209 597
25	1.182 478 058(L)	27	0.022 553 238
4	1.090 893 523(L)	7	0.022 206 826
7	1.081 879 146(L)	8	0.021 204 055

为便于作图比较，将 KRI 指数取对数，KRI 值大于 100 的标注(H)，KRI 值小于 20 的标注(L)。

4.6 KRI 指标对最优主题数目确定的指导意义

在 LDA 模型中，主题数目 K 的取值直接影响到模型质量和主题生成，主题数目过多或过少都会对主题识别结果产生影响。图 1 困惑度－主题相似度曲线中，K＝30 和 K＝45 时都出现拐点，表明在这两处 LDA 模型的泛化能力均显著增强。K＝45 之后，困惑度曲线趋于平缓，几乎不再下降，从这个角度看，K＝45 时主题识别效果更佳。但从主题平均相似度的角度看，K＝30 时平均 JS 距离更大，主题平均相似度更低(见表 4)，识别效果更好。那么，30 和 45 哪个是最优主题数目呢?

表 4 K＝30 和 K＝45 时困惑度、平均 JS 距离对比

主题数目	困惑度	平均 JS 距离
30	3 830.327	0.477
45	3 600.460	0.418

通过人工判读分析出 K＝30(见表 2)和 K＝45(见表 5)时主题识别结果并进行 KRI 指数计算及统计分析：

K＝45 时识别出的无效主题有 19 个，占 42.2%，远高于 K＝30 时的无效主题占比 20%(见表 6)。两种主题数目下，能识别出一些共同有效主题，如高 KRI 值的核心主题“微生物燃料电池”“单细胞凝胶电泳”“单细胞油脂”“胚胎植入前遗传学诊断”等。同时，两种主题数目下识别出非共同有效主题，经人工判读发现 K＝30 时主题有效程度更高。K＝45 时，主题“空气阴极微生物燃料电池”与“微生物燃料电池性能”均为“微生物燃料电池”的子主题，说明 K＝45 时主题识别粒度偏小。综合考虑可以得出，K＝30 时 LDA 模型的主题识别效果更好。由此可见，通过困惑度－主题相似度曲线选取最优主题数目时，可以计算不同拐点处主题数目下的 KRI 指数，比较分析辅助判断最优主题数目。

5 结语

LDA 模型识别结果中噪声主题的存在影响了主题识别和后续主题演化分析的准确性。本文提出一种基于关键词关联度指标(KRI)进行主题过滤的方法，能够有效过滤无确切含义的噪声主题，提高了主题识别结果的精准性，保证了后续主题演化路径构建的科学性。比较分析发现，KRI 指标方法的主题过滤效果优于“基于主题分布的边缘主题识别与过滤”方法。KRI 指标可以在一定程度上降低主题识别过程对人工判读的过度依赖，对于最优主题数目的选择也具有一定的参考作用。

表 5　K = 45 时 LDA 主题识别结果按 KRI 指数降序排列

KRI 指数	主题内容	KRI 指数	主题内容
2 011.294 889(H)	微生物燃料电池	21.572 499 19	胚胎发生
1 164.211 795(H)	单细胞凝胶电泳	20.814 238 04	遗传工程改造植物细胞壁
1 160.043 652(H)	单细胞油脂	20.510 889 63	单细胞分析方法
1 013.663 676(H)	胚胎植入前遗传学诊断	19.445 412 31(L)	无效主题
864.246 296 6(H)	空气阴极微生物燃料电池	18.136 209 81(L)	无效主题
435.073 998 3(H)	胞内钙调控	17.140 422 63(L)	无效主题
350.816 584 4(H)	干细胞体外培养	16.811 428 57(L)	无效主题
161.199 070 7(H)	细胞表面互作成像	12.024 556 62(L)	无效主题
132.747 798 2(H)	肿瘤诊断和治疗	10.581 726 74(L)	无效主题
125.680 585(H)	运动的神经调节	10.258 317 03(L)	无效主题
116.797 505 5(H)	基因及蛋白表达调控	8.706 666 667(L)	无效主题
109.034 513 7(H)	细胞迁移	8.701 504 355(L)	无效主题
107.947 918 6(H)	免疫反应	8.252 786 221(L)	无效主题
82.021 390 37	群体异质性演化	7.819 672 131(L)	无效主题
60.358 184 76	微生物燃料电池性能	7.426 499 033(L)	无效主题
59.172 449 51	细胞凋亡和细胞坏死	7.073 170 732(L)	无效主题
55.217 066 67	细胞电生理	6.679 146 812(L)	无效主题
48.393 026 57	单细胞质谱	6.660 066 007(L)	无效主题
41.035 398 23	神经胶质瘤类型分析	6.531 147 541(L)	无效主题
34.809 663 87	生物降解	6.356 25(L)	无效主题
30.060 753 34	肿瘤异质性	6.269 982 238(L)	无效主题
22.500 496 52	胞外电子传递	5.983 132 53(L)	无效主题
22.418 181 82	蛋白质稳定性		

表 6　K = 30 和 K = 45 时 LDA 识别结果中有效主题及噪声主题占比分析

主题数目	有效主题数	无效主题数	无效主题占比
30	24	6	20%
45	26	19	42.2%

但是,KRI 指标在有效主题和无效主题之间没有明显清晰的界限,从有效主题到无效主题 KRI 指标不是“断崖式”下降,在进行主题过滤时只能起到参考作用,不是主题有效性的绝对判断指标,仍然需要结合人工判读做出最终的取舍。本研究只是验证了 KRI 的主题过滤方法对于 LDA 模型有效,是否适用于其他主题模型识别结果的主题过滤有待于进一步的研究。

参考文献:

[1] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of machine learning research, 2003(3): 993 - 1022.

[2] BLEI D M, LAFFERTY J D. Dynamic topic model[C]//Proceedings of the 23rd international conference on machine learning. New York: ACM, 2006:113 - 120.

[3] WANG X R, MCCALLUM A. Topic over time: A non-markov continuous-time model of topical trends[C]//Proceedings of the 12th ACM SIG KDD International conference on knowledge discovery and data mining. Philadelphia: ACM, 2006: 424 - 433.

[4] YAN X H, GUO J F, LAN Y Y, et al. A biterm topic model for short texts[C]//Proceedings of the 22nd international conference on World Wide Web. New York: ACM. 2013:1445 - 1455.

[5] ZHAO F, ZHU Y J, JIN H, et al. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment[J]. Future generation computer systems, 2016, 65: 196 - 206.

[6] MAGNUSSON M, JONSSON L, VILLANI M. DOLDA: a regularized supervised topic model for high-dimensional multi-class regression[EB/OL]. [2019 - 09 - 08]. https://doi.org/10.1007/s00180-019-00891-1.

[7] 解琰. 主题优化过滤方法与研究应用[D]. 大连:大连海事大学, 2015: 26 - 27.

[8] 曲佳彬,欧石燕. 基于主题过滤与主题关联的学科主题演化分析[J]. 数据分析与知识发现,2018,2(1): 64 - 75.

[9] MACKAY D J C. Information theory, inference, and learning algorithms[M]. Cambridge:Cambridge University Press, 2003.

[10] 李保利,杨星. 基于 LDA 模型和话题过滤的研究主题演化分析[J]. 小型微型计算机系统,2012,3(12): 2738 - 2743.

[11] ISHWARAN H, RAO J S. Spike and slab gene selection for multigroup microarray data[J]. Journal of the American Statistical As-

sociation, 2005, 100(471): 764－780.

[12] CHANG Y L, LEE K F, CHIEN J T. Bayesian feature selection for sparse topic model[C]//IEEE international workshop on machine learning for signal processing (MLSP). Santander:IEEE, 2011:1－6.

[13] PONWEISER M, GRUN B. Finding scientific topics revisited [C]//CARPITA M, BRENTARI E, QANNARI E M. Advances in latent variables. Berlin: Springer, 2014: 93－100.

[14] 关鹏,王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术, 2016(9):42－50.

[15] GROSSMAN D A, Frieder O. Information retrieval: algorithms and heuristics[M]. Berlin: Springer, 2004.

[16] LEE L. On the Eectiveness of the skew divergence for statistical language analysis[C]//RICHARDSON T S, JAAKKOLA T S. Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics. Key West: Society for Artificial Intelligence and Statistics, 2001: 65－72.

[17] CAO J, XIA T, LI J, et al. A density-based method for adaptive LDA model selection[J]. Neurocomputing, 2009, 72(7/9): 1775－1781.

[18] CALLON M, COOUTIAL J P, LAVILLE F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry[J]. Scientometrics, 1991, 22(1): 155－205.

[19] WANG Z Y, LI G, LI C Y, et al. Research on the semantic-based co-word analysis[J]. Scientometrics, 2012, 90(3): 855－875.

[20] TURNER K, LYNCH C, ROUSE H, et al. Direct single-cell analysis of human polar bodies and cleavage-stage embryos reveals no evidence of the telomere theory of reproductive ageing in relation to aneuploidy generation[J]. Cells, 2019,8(2):1－17.

[21] FLETCHER R B, DAS D, GADYE L, et al. Deconstructing olfactory stem cell trajectories at single-cell resolution[J]. Cell stem cell, 2017, 20(6):817－830.

[22] JACOBSEN S E W, NERLOV C. Haematopoiesis in the era of advanced single-cell technologies[J]. Natrue cell biology, 2019, 21(1):2－8.

[23] GERDES M J, GÖKMEN-POLAR Y, SUI Y, et al. Single cell heterogeneity in ductal carcinoma in situ of breast[J]. Modern pathology, 2018,31(3):406－417.

[24] DAVIS K M, ISBERG R R. Defining heterogeneity within bacterial populations via single cell approaches[J]. Bioessays, 2016, 38(8):782－790.

[25] KOSTOFF R N. Co-word analysis[C]//BOZEMAN B, MELKERS J. Evaluating R&D impacts: methods and practice. New York: Springer, 1993:63－78.

作者贡献说明：

蒋甜：提出研究思路及技术路线，进行实验，分析数据，论文撰写；

刘小平：论文修改；

刘会洲：提出论文研究方向，论文修改及最终版本修订。

Topic Filtering of LDA Model Recognition Results Based on the Keywords Relevance Index (KRI)

Jiang Tian　Liu Xiaoping　Liu Huizhou
National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] The identification results of the LDA model is sometimes unsatisfactory due to some meaningless topics mixed together. Therefore, it's quite necessary to establish an effective topic filtering method to eliminate these noise topics and to ensure the accuracy of subsequent evolution analysis. [**Method/process**] Based on the co-occurrence relationship between keywords, keywords relevance index (KRI) was constructed. Taking the field of single cell research as an example, KRI values of the distribution of theme-keywords were calculated and compared with the results of manual interpretation. [**Result/conclusion**] Experimental results show that this method can effectively eliminate meaningless noise topics in the LDA model recognition results, which can improve the accuracy of topic recognition and the subsequent topic evolution analysis. It also helps to reduce the dependence on manual interpretation in the process of topic identification through the topic model method.

Keywords: topic filtering　LDA model　keywords relevance index (KRI)